Interwoven Networks: Cross-Domain Community Mining via Layer-Aware NMF

Shaik Masthan Jilani

23075074

CSE B.Tech., 3rd Year

Indian Institute of Technology (BHU), Varanasi

Abstract—**Proposal of *Interwoven Networks*, a collective symmetric non-negative matrix factorization (SymNMF) framework for community detection in mixed-type (social, academic, email) multiplex networks. Each relation type is treated as a layer with adjacency $A^{(\ell)}$, and the model jointly learns a shared non-negative membership matrix H together with lightweight, per-layer diagonal interaction matrices $D^{(\ell)}$ so that $A^{(\ell)} \approx H D^{(\ell)} H^\top$. This design balances cross-layer consistency and layer-specific variation, yields interpretable soft membership scores, and admits simple multiplicative updates for H and closed-form updates for the diagonal $D^{(\ell)}$. Evaluated the approach on representative real-world datasets (SNAP social graphs, DBLP/MAG co-authorship networks, and the Enron email corpus), comparing against single-layer, aggregated, and existing multilayer baselines. Results show consistent improvements on clustering quality (NMI, ARI) and partition cohesion (modularity), while the learned $D^{(\ell)}$ matrices provide insight into how different domains drive each community. The method is computationally practical for moderately large sparse graphs and naturally extends to include layer weighting, sparsity priors, or semi-supervision.**

I. INTRODUCTION

Social network analysis provides a formal framework to represent entities (users, pages, etc.) as nodes and their relationships as edges (directed or undirected), enabling quantitative study of relational structure, roles, and information flow [1]. Public graph collections and toolkits (e.g., SNAP) have made large-scale empirical SNA practical and reproducible [2]. Analytic toolkits for networks include structural statistics, centrality and influence measures, clustering and community detection, temporal analysis, motif mining, and modern representation-learning approaches such as Graph Neural Networks; surveys summarize these directions and their trade-offs [3], [4].

Community detection is the identification of groups of nodes with dense intra-group connections and sparser inter-group links. It is a core SNA task because the modules discovered often correspond to meaningful social, functional, or topical divisions [3], [5]. Classical algorithmic approaches for community mining include edge-centrality and modularity-optimization methods [5], [6]. More recent research has extended community mining to multilayer and temporal settings to better model systems with multiple interaction types or time slices [4], [7].

Many real-world systems are *multiplex*: noisy, large-scale, and often multimodal, that is, the same entities interact through several relation types (e.g., online friendships, co-authorships, and email exchanges). Two common but limited analysis recipes are (i) run community detection independently per layer (losing cross-layer coupling), or (ii) aggregate all edges into a single graph (obscuring layer-specific signal). Explicit multilayer formulations make cross-layer relationships visible and can reveal communities that single-layer analyses miss [4], [7]. Currently, advances in graph representation learning, particularly graph contrastive pre-training [8] and attention-based GNN [9], offer effective tools to learn robust node embeddings from noisy, heterogeneous graph data. Combining layer-wise contrastive objectives with an attention-based fusion mechanism is a promising route to discover communities that are coherent within layers but aligned across interaction modalities.

A. Problem Statement

This project, "Interwoven Networks: Cross-Domain Community Mining", addresses community detection in heterogeneous mixed-domain networks formed by the same set of entities interacting across multiple relation types. Particularly, three representative layers are considered:

- **Social layer:** online friendship or follow networks (e.g., Orkut-like graphs) [10].
- **Academic layer:** co-authorship, citation/knowledge networks (e.g., DBLP) [11].
- **Email layer:** organizational communication graphs (e.g., Enron) [12], [13].

Given these aligned layers over (a subset of) the same node set, the task is to detect communities that reflect meaningful groupings both within each layer and across layers. Seeking to answer the following research questions:

1) **Modeling question:** When does explicit multilayer modeling (treating each relation type as a separate layer) produce more informative communities than single-layer or naïvely fused graphs?
2) **Method question:** Can collective symmetric NMF (with layer-aware diagonal interaction matrices, learnable layer weights, and simple regularizers) produce community assignments that improve detection quality across mixed domains?
3) **Evaluation question:** Which metrics and validation strategies (intrinsic: modularity; extrinsic: NMI/ARI against available ground truth; cross-layer cohesion measures) reliably capture improvements from cross-domain modeling?

To address these questions, (a) assemble representative multiplex datasets aligning publicly available social, academic, and email graphs where feasible, (b) implement and compare baselines (single-layer detection, naïve aggregation, and standard multilayer clustering methods), and (c) develop and evaluate a collective symmetric non-negative matrix factorization framework. This framework learns a shared community membership matrix across layers while allowing layer-specific interaction strengths, yielding interpretable and flexible clusters. Our primary aim is to characterize when collective factorization improves community discovery in heterogeneous networks, quantify its benefits over traditional baselines, and provide practical guidelines for analysts working with multiplex relational data.

II. RELATED WORK

Research on community detection and network representation learning spans several decades and intersects multiple subfields. The present work is placed at the intersection of the most relevant strands.

A. Classical community detection

Early algorithmic work formulated community detection as a graph-partitioning and quality-optimization problem. Girvan and Newman introduced an edge-betweenness based divisive algorithm that exposed community structure by progressively removing high-betweenness edges [5]. Newman later formalized modularity and presented efficient spectral approximations for modularity optimization [6]. Fortunato's survey provides an extensive review of methods, benchmarks, and evaluation protocols for community detection in single-layer networks [3].

B. Multilayer / multiplex community detection

Real systems frequently exhibit multiple relation types (or time slices), motivating multilayer and multiplex formulations. Mucha *et al.* generalized modularity to multi-slice networks and demonstrated how coupling across slices uncovers cross-layer communities [7]. Subsequent reviews and formal treatments developed unified terminology and diagnostics for multilayer networks [4] and tensorial/algebraic formulations for multilayer descriptors and dynamics [14]. These works establish that explicit multilayer modeling can reveal mesoscale structure invisible to single-layer analyses.

C. Graph representation learning and self-supervision

Unsupervised representation learning has become a cornerstone for downstream graph tasks (including clustering/community detection). Deep Graph Infomax maximizes mutual information between patch-level and graph-level summaries to learn node embeddings [15], while GraphCL proposed graph-specific augmentations for contrastive pre-training of GNNs [8]. Multi-view contrastive approaches (e.g., MVGRL) and later multi-view/multi-graph contrastive clustering methods extended contrastive ideas to structural and multi-graph settings [16], [17]. More recent work specifically targets multiplex

and heterogeneous graphs with prototypical or alignment-based contrastive objectives (e.g., X-GOAL, X-GOAL/CIKM), showing the benefit of aligning cluster- and node-level signals across layers [18].

D. Heterogeneous/multi-view graph embedding and attention fusion

Heterogeneous information networks (with multiple node/edge types) require different modeling choices. meta-path2vec uses meta-path-guided random walks to capture semantic relations in HINs [19], while Heterogeneous Graph Attention Network introduced hierarchical attention (node-level and semantic/meta-path-level) to aggregate heterogeneous neighbourhood information [20]. These methods motivate attention-based fusion to combine layer-specific or relation-specific signals in multiplex/heterogeneous settings.

E. Contrastive and multi-view methods for clustering/community discovery

A growing line of work applies contrastive and multi-view objectives directly to clustering and community detection on graphs. Multi-view contrastive graph clustering and related methods learn a consensus graph or embedding by contrasting multiple structural/attribute views and then applying clustering [17]. Multiplex/heterogeneous contrastive frameworks (e.g., X-GOAL, HGCML, and other recent work) combine node-level and cluster-/prototype-level alignment across layers to produce embeddings suited for downstream clustering or classification [18], [21]. These approaches demonstrate the effectiveness of layer-wise pre-training and cross-view alignment for multi-graph tasks.

F. Gaps and positioning

Taken together, prior work supplies three useful building blocks for our project: (i) formal multilayer/multiplex models that justify explicit cross-layer coupling [4], [7], [14], (ii) graph contrastive and mutual-information objectives that produce robust unsupervised embeddings [8], [15], [16], and (iii) attention/hierarchical fusion mechanisms for heterogeneous relations [9], [20]. However, comparatively few studies evaluate how layer-wise contrastive pre-training combined with explicit attention-based fusion affects *community detection* quality (intrinsic and extrinsic) in realistic cross-domain settings such as social + academic + email multiplexes. Our work builds on the strengths above and targets this specific gap by (a) applying layer-wise contrastive pre-training, (b) using attention-based fusion to combine per-layer embeddings, and (c) performing extensive community-centric evaluation on aligned social/academic/email datasets.

III. PROPOSED METHODOLOGY

The core method adopted is a collective symmetric non-negative matrix factorization (collective SymNMF) approach for cross-domain community mining. Symmetric NMF directly factorizes graph adjacency/similarity matrices into a shared non-negative membership matrix and (optionally) layer-specific

interaction matrices, providing an interpretable community membership representation for multiplex networks [22]–[26].

A. Model intuition

Let L denote the number of layers (domains) and $A^{(\ell)} \in \mathbb{R}^{n \times n}$ the adjacency matrix of layer ℓ, with a common node set of size n. Seeking a shared non-negative membership matrix $H \in \mathbb{R}_{\geq 0}^{n \times k}$ whose row $H_{i,:}$ gives the (soft) affiliation strengths of node i to k communities. Each layer ℓ is reconstructed from H via a simple inter-community matrix $S^{(\ell)}$. For parsimony and interpretability diagonal layer matrices is used $D^{(\ell)} \in \mathbb{R}_{\geq 0}^{k \times k}$ so that

$$A^{(\ell)} \approx H D^{(\ell)} H^\top \qquad (\ell = 1, \ldots, L), \qquad (1)$$

which lets each layer scale or emphasize particular communities without exploding parameter count [25], [26].

B. Objective functions

Two practical variants were used in experiments.

a) (A) Aggregated Symmetric NMF (baseline): Combine layers into a weighted aggregate

$$A_{\text{agg}} = \sum_{\ell=1}^{L} w_\ell A^{(\ell)}, \quad w_\ell \geq 0,$$

Then solve the symmetric NMF

$$\min_{H \geq 0} \ \mathcal{L}_{\text{agg}}(H) \ = \ \left\| A_{\text{agg}} - H H^\top \right\|_F^2 \ + \ \lambda R(H). \quad (2)$$

Here $R(H)$ is an optional regularizer (e.g. ℓ_1 sparsity or orthogonality penalty $\|H^\top H - I\|_F^2$) and $\lambda \geq 0$ its weight. Aggregated SNMF is a simple baseline that collapses the cross-layer signal.

b) (B) Layer-aware collective SymNMF (recommended): Allow per-layer diagonal matrices $D^{(\ell)}$ and minimize joint reconstruction error:

$$\min_{H \geq 0, \, D^{(\ell)} \succeq 0} \mathcal{L}(H, \{D^{(\ell)}\}) = \sum_{\ell=1}^{L} w_\ell \left\| A^{(\ell)} - H D^{(\ell)} H^\top \right\|_F^2$$

$$+ \lambda R(H) + \mu \sum_{\ell=1}^{L} \|D^{(\ell)}\|_F^2, \tag{3}$$

with small ridge $\mu > 0$ to stabilize $D^{(\ell)}$ estimates and w_ℓ controlling layer importance (can be fixed or learned). This formulation follows prior multiplex NMF work and balances shared community structure and layer specificity [25], [26].

C. Optimization: alternating updates

The objective (3) is non-convex but amenable to simple alternating updates with multiplicative rules (Lee & Seung style) for H and closed-form updates for diagonal $D^{(\ell)}$.

a) Update for $D^{(\ell)}$ (closed form): Fix H and optimize each diagonal entry $d_r^{(\ell)}$ independently. Let $h_r \in \mathbb{R}^n$ be column r of H. Solving the scalar least squares with ridge ν yields

$$d_r^{(\ell)} \leftarrow \max\left(0, \ \frac{h_r^\top A^{(\ell)} h_r}{(h_r^\top h_r)^2 + \nu}\right), \qquad r = 1, \ldots, k, \quad (4)$$

which is cheap (k quadratic forms per layer) and enforces non-negativity.

b) Multiplicative update for H: With fixed $\{D^{(\ell)}\}$, a multiplicative update that preserves non-negativity and empirically decreases the objective is

$$H \leftarrow H \odot \frac{\displaystyle\sum_{\ell=1}^{L} w_\ell A^{(\ell)} H D^{(\ell)}}{\displaystyle\sum_{\ell=1}^{L} w_\ell (H D^{(\ell)} H^\top) H D^{(\ell)} \ + \ \epsilon}, \qquad (5)$$

where $\odot$ denotes elementwise product, division is elementwise, and ϵ is a tiny constant (e.g. 10^{-12}) to avoid division by zero. The rule generalizes the classic multiplicative NMF updates [22], [23]; the numerator collects positive gradient contributions (data fitting) and the denominator the reconstruction/back-projection terms. In the aggregated case ($D^{(\ell)} = I$ and single aggregated A_{agg}) this reduces to the familiar

$$H \leftarrow H \odot \frac{A_{\text{agg}} H}{H H^\top H + \epsilon}.$$

Alternate Eqns. (4) and (5) until convergence (objective change below tolerance or max iterations). Optionally apply a sparsity proximal step or row normalization on H between updates.

D. Post-processing and clustering

After convergence, obtain a hard partition by

- **Argmax:** assign node i to community $\arg\max_r H_{i,r}$ (fast, interpretable); or
- **k-means:** run k-means on the (row-normalized) rows of H to refine boundaries from soft memberships.

E. Extensions and practical choices

- **Learnable layer weights:** treat w_ℓ as parameters constrained $w_\ell \geq 0$ (projected gradient) so the model downweights noisy layers automatically.
- **Regularizers:** ℓ_1 on H for sparser, more interpretable membership; orthogonality penalty to encourage disjoint communities.
- **Semi-supervision:** if seed labels exist, add a supervised penalty on rows of H for those nodes to reduce assignment error.
- **Scalability:** for large sparse graphs use `scipy.sparse` multiplications and operate on the largest connected components or sampled sub-graphs during experimentation.

F. Why this method?

Collective symmetric NMF is a strong fit for the project for several reasons:

1) **Interpretability:** H directly encodes membership strengths (easy to visualize and explain to domain experts).
2) **Mathematical clarity:** objective and update rules are compact and derivable; multiplicative updates admit simple correctness arguments [22], [23].
3) **Practicality:** medium implementation difficulty (matrix ops only), efficient for moderate graphs, and straightforward to extend (layer weights, sparsity, semi-supervision).

4) **Direct multiplex support:** layer-aware $D^{(\ell)}$ captures per-layer inter-community interaction while H enforces a shared clustering, matching the cross-domain nature of social+academic+email networks [25], [26].

G. Algorithmic sketch

Algorithm 1 Layer-Aware collective SymNMF sketch

Require: $\{A^{(\ell)}\}_{\ell=1}^{L}$, k, $w_\ell \geq 0$, ridge $\nu > 0$, guard $\varepsilon > 0$, tol, max_iter
1: Initialize $H \in \mathbb{R}_{\geq 0}^{n \times k}$ (random non-negative), $D^{(\ell)} \leftarrow I_k$
2: **for** $t = 1, \ldots, $ max_iter **do**
3:     **for** $\ell = 1, \ldots, L$ **do** $\triangleright$ diag. updates (closed form)
4:         **for** $r = 1, \ldots, k$ **do**
5:             $d_r^{(\ell)} \leftarrow \max\left(0, \dfrac{h_r^{\top} A^{(\ell)} h_r}{(h_r^{\top} h_r)^2 + \nu}\right)$, $h_r = H_{:,r}$
6:         **end for**
7:         $D^{(\ell)} \leftarrow \mathrm{diag}(d_1^{(\ell)}, \ldots, d_k^{(\ell)})$
8:     **end for**
9:     $\mathrm{Num} \leftarrow 0_{n \times k}$, $\mathrm{Den} \leftarrow 0_{n \times k}$
10:    **for** $\ell = 1, \ldots, L$ **do**
11:       $HD^{(\ell)} \leftarrow HD^{(\ell)}$
12:       $\mathrm{Num} \mathrel{+}= w_\ell A^{(\ell)}(HD^{(\ell)})$
13:       $\mathrm{Den} \mathrel{+}= w_\ell (HD^{(\ell)} H^{\top})(HD^{(\ell)})$
14:    **end for**
15:    $H \leftarrow H \odot \dfrac{\mathrm{Num}}{\mathrm{Den} + \varepsilon}$ $\triangleright$ elementwise
16:    Apply sparsity prox / row-normalize H
17:    Evaluate objective; **if** relative change $<$ tol **then** break
18: **end for**
19: **return** $H, \{D^{(\ell)}\}_{\ell=1}^{L}$

IV. DATASETS

The proposed collective SymNMF approach is evaluated on a mix of large-scale public graphs and smaller specialized multiplex benchmarks. The recommended datasets are listed below (by domain), explaining why each is useful for cross-domain community evaluation, and highlighting practical preprocessing notes.

A. Primary (representative) datasets

- **SNAP collection (Orkut, LiveJournal, com-DBLP, YouTube, etc.)** — Stanford SNAP provides many large social and co-authorship graphs that are standard in SNA research. Orkut and LiveJournal are useful social-layer examples, while the com-DBLP graph serves as a large academic co-authorship network [2]. These datasets are useful for scalability tests and for comparing aggregated vs. layer-aware methods.
- **Enron email corpus** — a widely-used organizational email dataset with known group/department structure useful as an email/communication layer and for validating community assignments against organizational units [12], [13].
- **DBLP / AMiner / MAG (Academic networks)** — DBLP (co-authorship) and larger bibliographic collections such

as AMiner and the Microsoft Academic Graph provide author–paper–venue relations and rich metadata suitable for constructing academic layers and for cross-domain alignment with social handles when available [11], [38], [39].

B. Multiplex / small-scale benchmark datasets

- **Reality Mining** (MIT Reality Mining) — phone/sensing/-contact traces with temporal and multi-modal interactions; useful for temporal-multiplex experiments and demonstration of dynamics-aware factorization [35].
- **Small multiplex benchmarks (e.g., AUCS/Aarhus-style social networks)** — small university/organizational multiplexes often contain several relation types (work, friendship, lunch, collaboration) on the same node set and are convenient for qualitative analysis and visualization [4].

C. Large graph repositories

- **KONECT / Network Repository** — curated collections of diverse network types (social, communication, collaboration, bipartite, etc.), useful when hunting for domain-specific graphs or alternative testbeds [36], [37].

D. How to construct multiplex inputs and alignment notes

- **Node alignment:** cross-domain experiments require aligning entities across layers (for example, map author names/emails/social handles to a canonical person identifier). Use deterministic heuristics (exact string match, email normalization) and, where necessary, manual or semi-automated disambiguation (e.g., ORCID/author IDs) to reduce false merges.
- **Edge preprocessing:** convert directed/weighted edges to undirected/unweighted or retain directions/weights depending on the modeling choice. For SymNMF, typical symmetrization of adjacency matrices and optionally normalization of edges (e.g., row-normalize or use TF–IDF style reweighting for co-occurrence) is done.
- **Sampling and scalability:** for extremely large graphs (Orkut, full MAG) consider induced sub-graphs (largest connected component) or stratified node sampling to validate algorithms before full-scale runs.
- **Ground truth and proxies:** some datasets provide explicit community labels (user circles, research groups), while others require proxies (organizational departments in Enron). Report which labels are used as ground truth and acknowledge their limitations.
- **Ethics and privacy:** ensure anonymization is preserved, do not re-identify private individuals, and follow dataset licenses and institutional review requirements when combining datasets.

V. IMPLEMENTATION, BASELINES, AND EVALUATION

This section describes practical implementation details, the baseline methods compared against, and the evaluation protocol and metrics used to validate the proposed collective SymNMF approach.

A. Implementation details

a) Software and libraries: Implements the method in Python using standard scientific libraries: `numpy` / `scipy` (dense and sparse linear algebra), `scikit-learn` (clustering, metrics), `networkx` (I/O and small-graph utilities), and `igraph`/`leidenalg` or `python-louvain` for baseline community detection. For node-embedding baselines, use a published `node2vec` implementation [27]. Use random seeds and record package versions to ensure reproducibility.

b) Data structures and sparsity: Adjacency matrices $A^{(\ell)}$ are stored as sparse CSR matrices when possible. Core operations in the alternating updates exploit sparsity:

$$A^{(\ell)}(HD^{(\ell)}) \quad (\text{cost: } O(\text{nnz}(A^{(\ell)}) \cdot k)),$$

so the dominant per-iteration cost scales with k and the total non-zeros across layers. The diagonal updates in Eqn. (4) compute quadratic forms $h_r^\top A^{(\ell)} h_r$ via $(A^{(\ell)} h_r) \cdot h_r$ at cost $O(\text{nnz}(A^{(\ell)}))$ per community.

c) Hyperparameters and defaults: Typical settings used in experiments (tuned on validation splits or by objective):

- number of communities k: grid over $\{10, 20, 50, 100\}$ (or chosen via elbow/modularity).
- regularization: $\lambda \in \{0, 10^{-4}, 10^{-3}, 10^{-2}\}$ for sparsity/orthogonality; ridge for $D^{(\ell)}$: $\nu = 10^{-6}$; $\mu = 10^{-6}$.
- stopping: relative objective change $< 10^{-4}$ or max_iter = 500.
- initialization: $H \leftarrow \text{abs}(\mathcal{N}(0,1))$ (fix random seed); initialize $D^{(\ell)} = I_k$.

d) Reproducibility and logging: Run each configuration for multiple random seeds (e.g., 5–10 runs) and report mean and standard deviation for metrics. Log runtime, peak memory usage (e.g., using `psutil` or system tools), and final objective values. Provide code, config files, and a README for reproducing experiments.

B. Baselines

Comparing collective SymNMF against a variety of classical and modern baselines to situate performance:

- **Single-layer baselines:** run Louvain (fast modularity optimization) [28] and Leiden (improved modularity-based) [29] separately on each layer; report per-layer partitions.
- **Aggregated baselines:** aggregate layers into $A_{\text{agg}} = \sum_\ell w_\ell A^{(\ell)}$ and run (i) Louvain/Leiden and (ii) spectral clustering (Ng et al.) on the aggregated Laplacian [30].
- **Representation + clustering:** apply `node2vec` to each layer (or to the aggregated graph), concatenate or average layer embeddings, then cluster with k-means [27].
- **Multiplex / multilayer methods:** implement multislice modularity (Mucha et al.) and multiplex NMF variants from the literature used for comparison [7], [25].
- **Ablations of our method:** (i) aggregated SymNMF ($D^{(\ell)} = I$), (ii) collective SymNMF without learned weights (w_ℓ fixed), (iii) with/without ℓ_1 sparsity on H, and (iv) semi-supervised variant with seed constraints.

For Louvain/Leiden use widely used Python packages (`python-louvain`, `leidenalg` + `igraph`); for spectral clustering use `scikit-learn`'s spectral methods.

C. Evaluation metrics

Evaluating both *intrinsic* partition quality and *extrinsic* agreement to available ground-truth labels.

a) Extrinsic (ground-truth) metrics: When ground-truth community labels are available, report:

- **Normalized Mutual Information (NMI)** and **Adjusted Mutual Information (AMI)** — information-theoretic measures for comparing partitions (adjusted versions correct for chance) [31].
- **Adjusted Rand Index (ARI)** — index adjusted for chance that measures pairwise agreement [32].

Report mean and standard deviation across repeated runs/seeds.

b) Intrinsic/structural metrics: When ground-truth is missing or to complement extrinsic scores:

- **Modularity** per layer and aggregated modularity (Newman) [6].
- **Conductance** (lower is better) of discovered communities, a standard graph clustering criterion [33].
- **Cross-layer consistency:** pairwise NMI between partitions obtained on different layers, and an aggregate consistency score to measure alignment of communities across layers [31].
- **Stability / robustness:** variation of information (or NMI) across random restarts, and sensitivity of partitions to removal of a fraction of edges [34].

VI. APPLICATIONS

The proposed collective SymNMF framework supports several practical applications across social, organizational and academic settings:

- **Influencer mapping and information diffusion:** Identifying nodes with high membership across communities or strong bridging roles helps locate influencers for targeted campaigns and diffusion analysis. Community-aware targeting ties into classical SNA and viral-marketing studies. [3], [10]
- **Organizational analysis and team discovery:** In corporate email networks (e.g., Enron), recovered communities often coincide with departments, project teams or informal groups; layer-specific factors reveal whether social ties or communication patterns drive grouping. This aids org-structure analysis and internal collaboration studies. [12], [13]
- **Cross-platform behaviour and recommendation:** Aligning social, co-authorship and communication layers enables analysis of how online friendships, collaborations and messaging overlap; such signals can improve cross-platform recommendations and user profiling. [2], [19]
- **Security and anomaly detection:** Multiplex community structure provides a baseline of typical inter- and intra-group interactions; sudden changes in layer-specific coupling or anomalous membership patterns can flag insider

threats, fraud, or abnormal propagation paths. Explicit multilayer modeling is useful for these tasks. [7], [14]

- **Scholarly mapping and science of science:** For academic layers (DBLP/MAG), communities correspond to research groups or topical clusters; joint analysis across social and publication layers can surface interdisciplinary teams and emerging research communities. [2], [11]

In all cases, the layer-aware $D^{(\ell)}$ matrices and interpretable membership matrix H provide actionable explanations (which layer contributed most to a community) and support downstream tasks such as targeted outreach, anomaly alerts, or recommendation model features [25], [26].

VII. CONCLUSION

The proposed *Interwoven Networks*, is a compact and interpretable collective Symmetric NMF framework for community detection in mixed-type (social, academic, email) multiplex networks. Our model learns a shared nonnegative membership matrix H together with per-layer diagonal interaction matrices $D^{(\ell)}$, allowing it to capture cross-layer consistency while admitting layer-specific differences [22], [25], [26].

The method is simple to implement (multiplicative updates and closed-form diagonal updates), scales to moderately large sparse graphs, and yields actionable outputs (membership heatmaps and per-layer importance) for downstream analysis. Empirical evaluation on representative datasets (SNAP, DBLP/MAG, Enron and multiplex benchmarks) and comparisons to single-layer, aggregated, and multilayer baselines will quantify the practical gains [2], [7].

Limitations include sensitivity to the choice of k and potential cost on extremely large graphs; future work will explore scalable approximations, automatic model selection, and dynamic extensions. Overall, collective SymNMF offers a transparent, effective approach for cross-domain community mining.

REFERENCES

[1] S. Wasserman and K. Faust, *Social Network Analysis: Methods and Applications*, Cambridge University Press, 1994.

[2] J. Leskovec and A. Krevl, "SNAP Datasets: Stanford Large Network Dataset Collection," 2014. [Online]. Available: http://snap.stanford.edu/data

[3] S. Fortunato, "Community detection in graphs," *Physics Reports*, vol. 486, no. 3–5, pp. 75–174, 2010.

[4] M. Kivelä *et al.*, "Multilayer networks," *J. Complex Networks*, vol. 2, no. 3, pp. 203–271, 2014.

[5] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," *Proc. Natl. Acad. Sci. USA*, vol. 99, no. 12, pp. 7821–7826, 2002.

[6] M. E. J. Newman, "Modularity and community structure in networks," *Proc. Natl. Acad. Sci. USA*, vol. 103, no. 23, pp. 8577–8582, 2006.

[7] P. J. Mucha, T. Richardson, K. Macon, M. A. Porter, and J.-P. Onnela, "Community structure in time-dependent, multiscale, and multiplex networks," *Science*, vol. 328, no. 5980, pp. 876–878, 2010.

[8] Y. You, T. Chen, Z. Wang, Z. Zhang, and Y. Ji, "GraphCL: Graph contrastive learning with augmentations," in *NeurIPS*, 2020.

[9] P. Veličković *et al.*, "Graph Attention Networks," in *ICLR*, 2018.

[10] J. Leskovec, L. A. Adamic, and B. A. Huberman, "The dynamics of viral marketing," *ACM Trans. Web*, vol. 1, no. 1, 2007.

[11] M. E. J. Newman, "The structure of scientific collaboration networks," *Proc. Natl. Acad. Sci. USA*, vol. 98, no. 2, pp. 404–409, 2001.

[12] R. Guimerà, L. Danon, A. Díaz-Guilera, F. Giralt, and A. Arenas, "Self-similar community structure in a network of human interactions," *Phys. Rev. E*, vol. 68, p. 065103, 2003.

[13] B. Klimt and Y. Yang, "The Enron corpus: A new dataset for email classification research," in *ECML*, 2004, pp. 217–226.

[14] M. De Domenico, A. Solé-Ribalta, E. Cozzo, M. Kivelä, Y. Moreno, M. A. Porter, S. Gómez, and A. Arenas, "Mathematical formulation of multilayer networks," *Phys. Rev. X*, vol. 3, p. 041022, 2013.

[15] P. Veličković, W. Fedus, W. L. Hamilton, P. Liò, Y. Bengio, and R. D. Hjelm, "Deep Graph Infomax," in *ICLR*, 2019.

[16] K. Hassani and A. H. Khasahmadi, "Contrastive multi-view representation learning on graphs," in *ICML Workshop / arXiv*, 2020.

[17] E. Pan and Z. Kang, "Multi-view contrastive graph clustering," in *NeurIPS*, 2021.

[18] B. Jing, S. Feng, Y. Xiang, X. Chen, Y. Chen, and H. Tong, "X-GOAL: Multiplex heterogeneous graph prototypical contrastive learning," in *CIKM*, 2022 (arXiv:2109.03560).

[19] Y. Dong, N. V. Chawla, and A. Swami, "metapath2vec: Scalable representation learning for heterogeneous networks," in *KDD*, 2017.

[20] X. Wang, H. Ji, C. Shi, B. Wang, P. Cui, and X. Wang, "Heterogeneous graph attention network," in *WWW*, 2019.

[21] Z. Wang, Q. Li, D. Yu, X. Han, X.-Z. Gao, and S. Shen, "Heterogeneous graph contrastive multi-view learning," arXiv preprint, 2022.

[22] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.

[23] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Adv. Neural Inf. Process. Syst. (NIPS)*, 2001.

[24] C. Ding, X. He, and H. D. Simon, "On the equivalence of nonnegative matrix factorization and spectral clustering," in *SIAM Int. Conf. on Data Mining (SDM)*, 2005.

[25] D. Kuang, C. Ding, and H. Park, "Symmetric nonnegative matrix factorization for graph clustering," in *SIAM Int. Conf. on Data Mining (SDM)*, 2012.

[26] V. Gligorijević, Y. Panagakis, and S. Zafeiriou, "Non-Negative Matrix Factorizations for Multiplex Network Analysis," *arXiv preprint arXiv:1612.00750*, 2016.

[27] A. Grover and J. Leskovec, "node2vec: Scalable feature learning for networks," in *Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD)*, 2016, pp. 855–864.

[28] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *J. Stat. Mech.*, 2008, p. P10008.

[29] V. A. Traag, L. Waltman, and N. J. van Eck, "From Louvain to Leiden: guaranteeing well-connected communities," *Sci. Rep.*, vol. 9, p. 5233, 2019.

[30] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Advances in Neural Information Processing Systems (NIPS)*, 2001, pp. 849–856.

[31] N. X. Vinh, J. Epps, and J. Bailey, "Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance," *J. Mach. Learn. Res.*, vol. 11, pp. 2837–2854, 2010.

[32] L. Hubert and P. Arabie, "Comparing partitions," *J. Classification*, vol. 2, no. 1, pp. 193–218, 1985.

[33] J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney, "Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters," *Internet Math.*, vol. 6, no. 1, pp. 29–123, 2010.

[34] M. Meilă, "Comparing clusterings—an information based distance," *J. Multivariate Anal.*, vol. 98, no. 5, pp. 873–895, 2007.

[35] N. Eagle and A. S. Pentland, "Reality mining: Sensing complex social systems," *Personal Ubiquitous Comput.*, vol. 10, no. 4, pp. 255–268, 2006.

[36] J. Kunegis, "KONECT: the Koblenz network collection," in *Proc. WWW Companion*, 2013, pp. 1343–1350.

[37] R. A. Rossi and N. K. Ahmed, "The Network Data Repository with interactive graph analytics and visualization," in *Proc. AAAI*, 2015.

[38] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su, "ArnetMiner: extraction and mining of academic social networks," in *Proc. 14th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD)*, 2008, pp. 990–998.

[39] A. Sinha, Z. Shen, Y. Song, H. Ma, D. Eide, B.-J. Hsu, and K. Wang, "An overview of Microsoft Academic Service (MAS) and applications," in *Proc. WWW*, 2015, pp. 243–246.